Classification and indexing of complex digital objects with CIDOC CRM

Juergen Enge; Head of Center for Media, Information & Technology at HAWK - University of Applied Sciences and Arts; Hildesheim, Germany. Tabea Lurk; Head of ArtLab – Conservation & Restoration at BUA - Bern University of Applied Sciences; Bern, Switzerland.

Abstract

CIDOC-CRM provides an ontology-based description for the documentation of cultural heritage. Originally meant to support the documentation practice of cultural heritage institutions and to enable inter-institutional exchange, it defines a formal structure for the description of implicit and explicit relations between entities.

*In order to demonstrate the benefits of the model in a semantic web environment like "Semantic MediaWiki", the paper shows two practical examples. Both originate in the digital domain and are complex due to their nature: As an example of a completely synthetically generated HD-Video, "Sintel" (2010) by Colin Levy is gathered. Facing distributed internet-based art and culture, Olia Lialinas "Summer" (2013) is described. The examples demonstrate in what extent the semantic structure of the digital extension of CIDOC CRM, which is CRM*dig, *clarifies the objects nature (understanding) and thus supports the planning and documentation process of dedicated collections. For doing so, an own system, called CRM-Wiki was implemented.*

Motivation

Archives and artworks in the cultural domain (excluding political archives and their objects) are very often diverse ([3]; [2]; [8]). They grow out of a broad history of collaboration and contributions of different types. Often they grow out of a broad history of collecting and storing activities, dedicated for collecting and contextualizing objects/information related to their host/museums, their collection(s), exhibition strategies, common policy or reference materials. Up to a certain point, these (file-) collections carry the thumbprint of different generations of scientists, archivists, curators, student assistants etc.

From documentation-based point of view analogies to the production of artworks or cultural productions can be seen. Here, too very often different persons, places, processes or resources are involved. One can say that especially in the digital domain, the production process of artworks or cultural goods in general carries the thumbprint of different materials/resources, people, places, software product and concepts. Beyond legal issues, a number of technological leaps are therefore challenging. Some of the information concerning the production or distribution process can be covered by metadata descriptions normally. As Doerr and Theodoridou (2013) state: "Generally, we use metadata to assess meaning (the recorded things, experimental setup, instrument used), relevance (status, conditions of the recording and derived information), quality (calibration, tolerances, measurement errors, processing artifacts and error propagation) and possibilities of improvement and data reprocessing." [4, p. 1] Nevertheless documenting the specific context of information is very often hard to track. Thus it seems reasonable to apply a rather abstract model than a fix metadata set in order to document complex, distributed artworks. The documentation process, as following exemplified, needs to be open and transparent by displaying the underlying semantics. Furthermore the system applied should be as simple to use and broadly known that, that the documentation is well understandable for people of different professions.

Methodological Approach

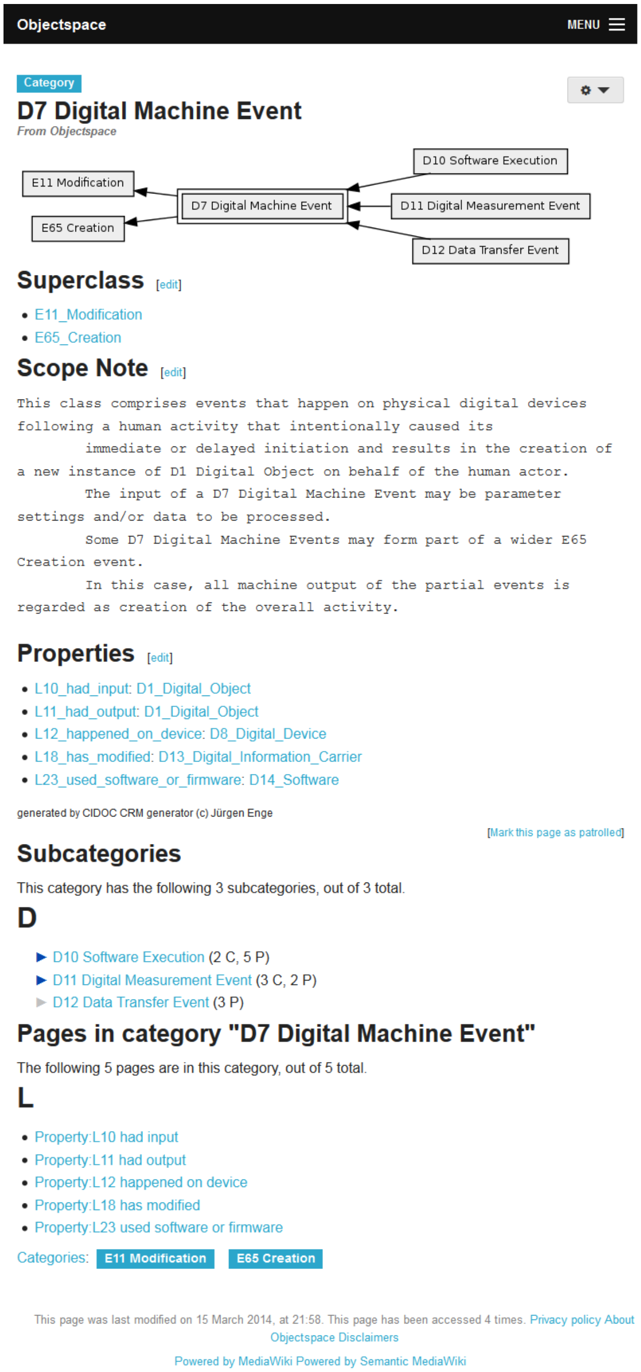
In order to raise the specific challenges occurring with complex born digital objects such as video and network based objects, we implemented the documentation concept of the ICOM standard CIDOC-CRM to a Semantic MediaWiki System. CIDOC CRM was released in 2003 and has been updated until lately [1]. Methodologically the model is a RDF schema, which consists of entities and properties. Both, entities and properties are organized in parent-child structures, whereas the properties define the specific relation semantically.

Beyond classical metadata sets, the broad set of semantic elements allows to document extensive production and realization processes. The mapping to accepted metadata schemas such as Dublin Core, OAIS, the CEDARS [2] or METS [3] has been proofed and can thus be considered accepted. Since our case studies are located in the digital domain, in addition to the original CIDOC-CRM-schema the digital extension CRM*dig* from 2013 is consulted [4].

CRM-Wiki

Initially, the CRM and CRM*dig* schemata were transferred to Semantic MediaWiki by inserting the entities and properties into the wiki system. Entities are represented in the Model with the prefix E-Entity (CIDOC) and D-digital Entity (CRM*dig*), whereas properties are represented by P-Property or L-digital Property. This enhanced Semantic MediaWiki system is called "CRM-Wiki" in the following.

Then an import-software was developed, which uses the CRM RDF schemata of definitions from CIDOC CRM and CRM*dig* as input for CRM-Wiki. This software maps the entities (E/D) to hierarchical wiki categories. "Categories are a means to classify articles according to certain criteria" [9, p. Editing]. The same is applied to the properties, which are mapped to native Semantic MediaWiki properties (P/L). Whereas the entities are derived from "superentities" and properties, properties are defined by their domain (as category) and range. In addition, properties can be derived from "superproperties". The CRM-Wiki thus generates automatically for each entity/property (of the CIDOC-CRM and the CRM*dig*) a wiki page that contains on one hand the rules, used to describe objects and on the other hand a full documentation of the entity/property. *Figure 1* shows as an example the CRM-Wiki-entry of the entity "D7 Digital Machine Event".



Objectspace MENU

Category

D7 Digital Machine Event

From Objectspace

E11 Modification

E65 Creation

D7 Digital Machine Event

D10 Software Execution

D11 Digital Measurement Event

D12 Data Transfer Event

Superclass [edit]

- E11_Modification
- E65_Creation

Scope Note [edit]

```
This class comprises events that happen on physical digital devices
following a human activity that intentionally caused its
        immediate or delayed initiation and results in the creation of
a new instance of D1 Digital Object on behalf of the human actor.
        The input of a D7 Digital Machine Event may be parameter
settings and/or data to be processed.
        Some D7 Digital Machine Events may form part of a wider E65
Creation event.
        In this case, all machine output of the partial events is
regarded as creation of the overall activity.
```

Properties [edit]

- L10_had_input: D1_Digital_Object
- L11_had_output: D1_Digital_Object
- L12_happened_on_device: D8_Digital_Device
- L18_has_modified: D13_Digital_Information_Carrier
- L23_used_software_or_firmware: D14_Software

generated by CIDOC CRM generator (c) Jürgen Enge

[Mark this page as patrolled]

Subcategories

This category has the following 3 subcategories, out of 3 total.

D

▶ D10 Software Execution (2 C, 5 P)
▶ D11 Digital Measurement Event (3 C, 2 P)
▶ D12 Data Transfer Event (3 P)

Pages in category "D7 Digital Machine Event"

The following 5 pages are in this category, out of 5 total.

L

- Property:L10 had input
- Property:L11 had output
- Property:L12 happened on device
- Property:L18 has modified
- Property:L23 used software or firmware

Categories: E11 Modification E65 Creation

This page was last modified on 15 March 2014, at 21:58. This page has been accessed 4 times. Privacy policy About Objectspace Disclaimers

Powered by MediaWiki Powered by Semantic MediaWiki

***Figure 1**. CRMdig Entity imported as MediaWiki category*

In addition to this ontological background information, which defines the semantic model, the browsing functionality of the Semantic MediaWiki allows to review dedicated semantic information (see *Figure 4*). If the syntax is applied in a proper way and as suggested below, an extensive explanation of each object is automatically generated and shown at each CRM-Wiki-page.

Furthermore the properties (references and literals) are added by inverse references, which are shown on the specific page (see below). Last but not least automatically generated graphs add a surplus-effect, since they make the specifically applied entities, properties and their relation among each other visually comprehensible (see below). The CRM-Wiki thus becomes a perfect tool for building up extensive CRM-compatible documentation systems, which is quite easy to use.

CRM-Wiki Syntax

The CRM-Wiki is meant to support planning processes in the development of new documentation systematics or within conceptual phases where collaboration with external archives or (meta-)data exports are considered. In order to generate meaningful entries, there is a need to use the predefined syntax of Semantic MediaWiki and the structure of CIDOC CRM / CRM*dig*. For demonstration purposes the following example might be supportive:

As a starting point a specific element/object is chosen – e.g. an artwork/person/event. The next step determines a mapping, where the objects/persons/events' nature is matched to a CRM entity. A wiki page is created with a unique name. This unique name should contain as a prefix the CRM entity number, e.g.:

"D9 Sintel Master".

D9 indicates the entity type "D9 Data Object". Assigning the page to the correct entity type is simplified by the category reference, which are integrated in the CRM-Wiki.

```
[[Category:D9 Data Object]]
```

In a similar way, properties are attached to the page (entity).

```
[[L58 has thumbnail::File:D39 Sintel.png]]<br />
Width: [[L56 has pixel width::4096]]<br />
Height: [[L57 has pixel height::1744]]<br />
Format: [[M1 codec::tiff16 stream]]<br />
Frames: [[M1 frames::21312]]
```

Again it seems important that the properties match with the definitions, given by the ICOM (CIDOC CRM, [1]) and FORTH group (CRM*dig* [5]). It's possible to add as much other wiki-content to the page as needed. However, this content is not part of the CRM structure. For understanding internal relations and the syntax, we suggest to follow the complete documentation. [9]

Within the documentation process evolving structures of properties and entities are "seamless" integrated in wiki pages by applying the correct code snippets. The CRM-Wiki (finally) frames the whole context of the documented content and relates to it additional information (see **Figure 2** and *Figure 3*).

In general properties direct from one entity to another by building up a directed semantic graph from property domain to range. For displaying inverse properties, which represent sort of backlinks, a helpful template using semantic queries is available. It can be integrated in the wiki pages by using the following code, whereas the second part is the property name, which should be used.

```
{{CRMInverse|P94 has created}}
```

The CRMInverse-template uses two semantic queries to discover the label of the inverse property and the domain of the property to display a correct reference.

```
{{#ask: [[{{{1}}}::{{FULLPAGENAME}}]]
| intro={{#ask: [[Inverseproperty
of::Property:{{{1}}}]]
| ?Has label=
| mainlabel=-
}}:&nbsp;
}}
```

Another helpful tool is the #sgraph command, which displays all page references as directed graph.

```
{{#sgraph:resource={{FULLPAGENAME}}
|depth=5 |engine=dot |svg=true |zoom=100% }}
```

Examination

In order to demonstrate the effect of the CRM-Wiki and to explain the shift from a relational, object-centered model to (action based) semantic web structures, a short comment on the two examples seems supportive.

Firtst Colin Levys film “Sintel” (2010) is introduced. Sintel, a small girl, is searching for a baby dragon, called Scales, in a dramatic scenic landscape. The film is mentioned to result of the film project “Durian”. It is one of the first HD creative common films. For us is important that many different people contributed to this film. Different versions in varying resolutions can be downloaded from the internet [6]. Both aspects are outlined in **Figure 2**.

As second example the network-based artwork “Summer” (2013) by Olia Lialina is documented [7]. The artist literally swings through the internet – entering page is

http://art.teleportacia.org/olia/summer/.

Produced as series of 21 image frames, on which the artists sits on a swing and swings continuously forward and backward, the challenge of the piece is that each frame is located on a different website. Looking at the URL-line of the browser it becomes obvious that the artwork is distributed over the internet. Identical frames are located at different servers around the world to minimize the risk of failure.

Whereas “Sintel” is based on a 3D-computeranimated model, “Summer” is a web-based.

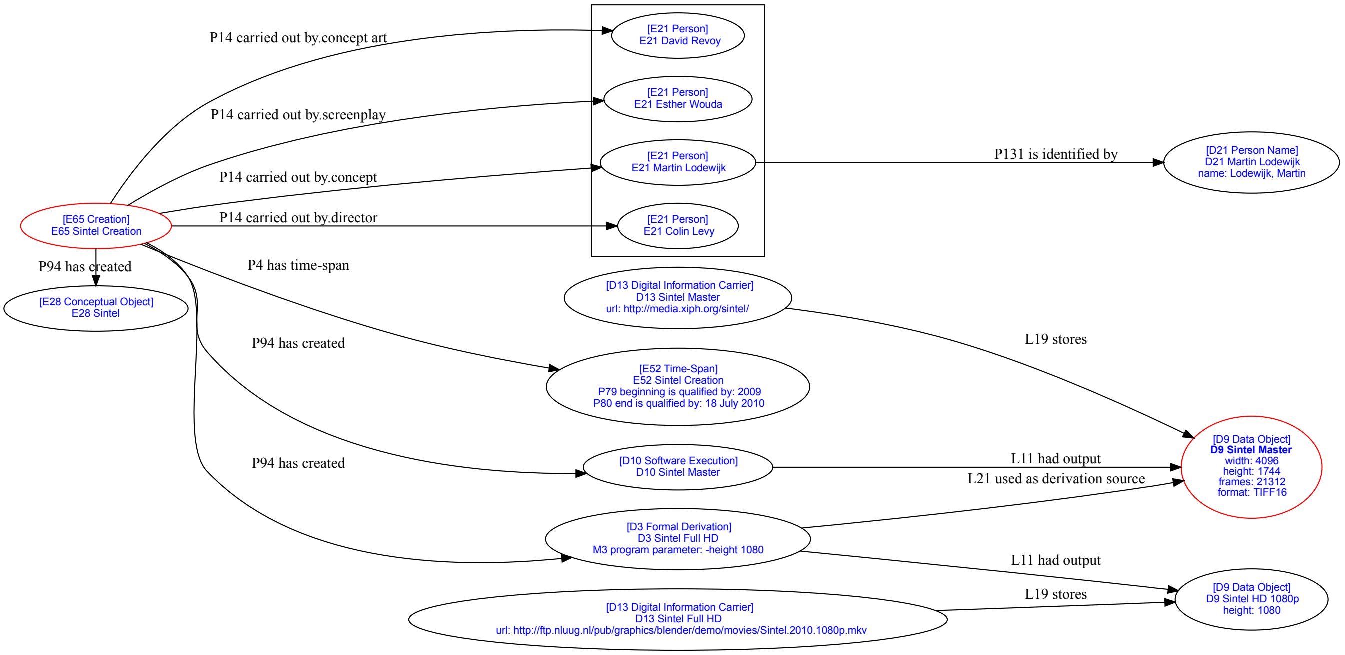


Figure 2. *Simplified CRM Graph of Sintel Documentation*

Description

Figure 2 shows an excerpt of the Sintel documentation graph. The center is the creation process [E65 Sintel Creation]. On the one hand this refers to the different people, which contributed to the film, like e.g. director Colin Levy (P14 carried out by:E21). On the other hand a conceptual object [E28 Sintel] is introduced (P94 has created), which frames the projects in its entirety.

Furthermore the software processes for the creation of the digital master ([D10 Sintel Master]), in terms of 3D-modelling and rendering, and the derivates ([D3 Sintel Full HD]) are documented. Additionally, metadata of the film-files and the information carrier ([D13]) for download is added to the system.

Figure 3 shows instead an excerpt of the Summer documentation graph. This time, besides the creation process [E65 Summer Creation] a software-driven “Digital Machine Event” [D7 Summer] is central, which can be understood as “events that happen on physical digital devices” [5, p. 6]. Furthermore the distributed character of the artwork is describes, which is performed in real time by accessing URLs on the internet. Since the frames are is located at different websited, these are addes as “Digital Information Carrier” [D13].

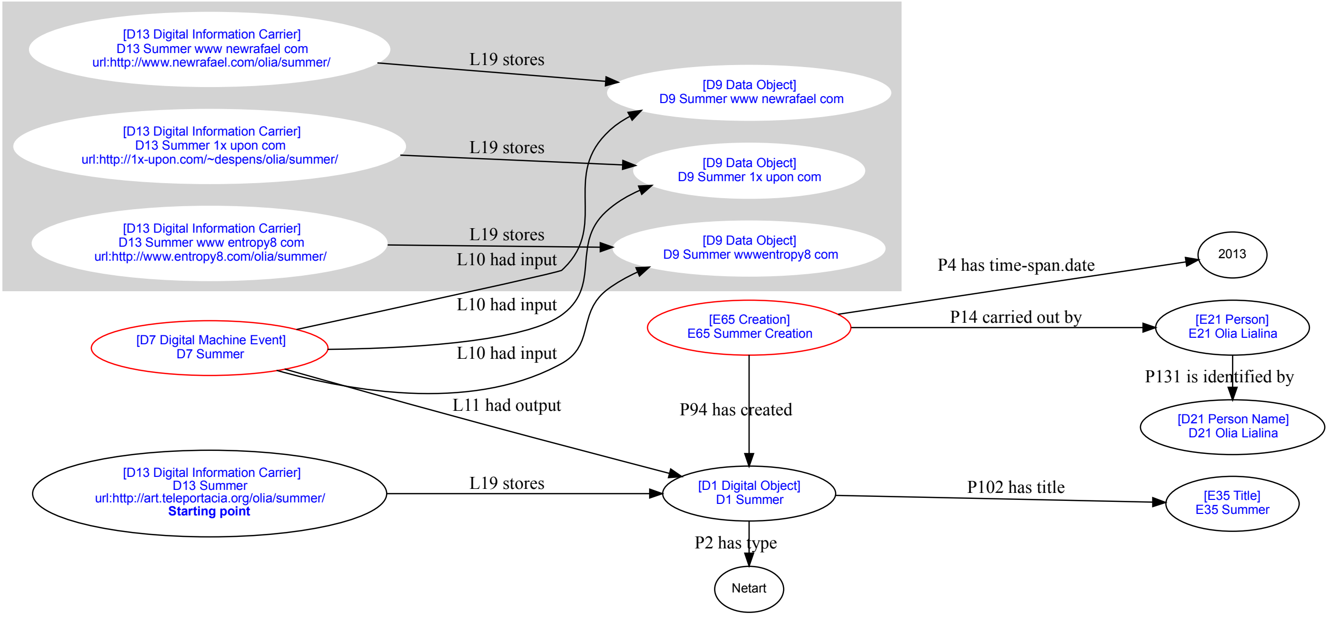


Figure 3. *Simplified CRM Graph of Summer*

Mapped to the structure of CRM*dig* one might say that [D7 Summer] consists of different “Data Objects” [D9], which reside on specific “Digital Information Carrier” [D13]. These data carrier are referenced by online addresses.

Results

The paper has shown that ICOMs CIDOC-CRM [1] and its digital extension CRM*dig* offer a process-centric approach. They have an abstraction level, which is capable of documenting not only static objects with a closed lifecycle, but also the requirements of continuously changing object types: The classical CIDOC-CRM has a very well defined set of entities with a focus on the area of activities [E7] e.g. for different types of actions (including the [E65 Creation], but also [E66 Formation], [E87 Curation Activities], as well as the [E8 Acquisition] event, [E10 Transfer of Custody]) [3]. The digital extension for example is capable of documenting a [D2 Digitization Process], including machinery chain [D8 Digital Device], digitization parameters (L10-L14) as well as process attribution (L21-24) [5]. Thus the object-centered cataloging of metadata could be abandoned in favor of a “Creation” [Event] (E65), which can be conducted by different persons, spread over time and space. Even the distributed performance of an event in real-time can be documented.

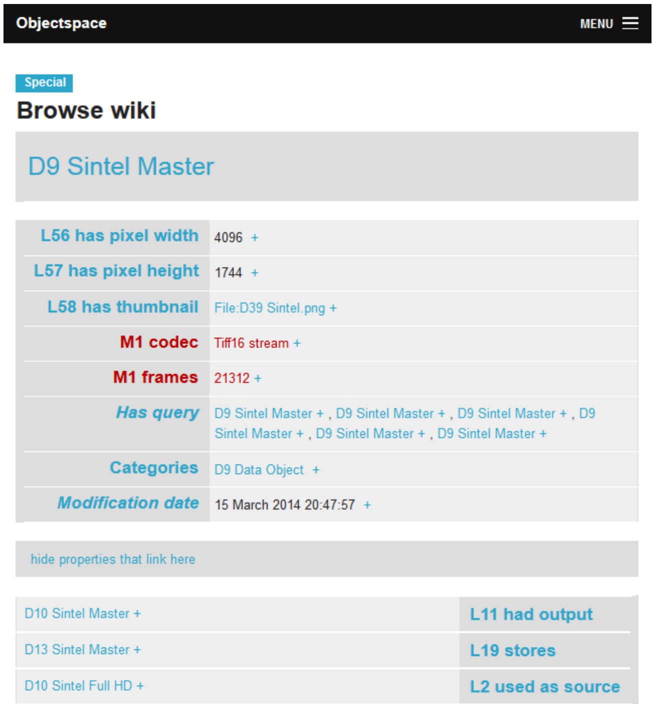
Whereas the digitization process of e.g. analog resources seems perfectly mapped/represented by the CRM*dig*, facing complex digital objects or net-phenomena limits become obvious. When implementing the CIDOC CRM, it is sometimes not possible to follow literally the predefined schema. In some cases, we enhanced the domain/range of some properties a bit to enable the description of unforeseen cases.

Another limit is given by the Semantic MediaWiki where properties can be cues as in the RDF-schema. Since CIDOC-CRM is implemented as RDF-technology, which is technologically more powerful than Semantic MediaWiki regarding the architecture of properties of properties, some restrictions occur. A CRM property can have properties by itself (i.e. roles of actors), whereas in the Semantic MediaWiki properties (edges) can’t be related to each other. In order to bypass this limitation we have introduced derived properties, which simulate this behavior.

Considering systematically applied semantic elements and tags instead, the system can guarantee that the basic knowledge of the systems is growing continuously. Controlled vocabularies, taxonomies and freely tagged information might be included in the future, in order to strengthen the internal ontology. The knowledge of the CRM-Wiki is not only based on the input information, data or other forms of “additional content”, it evolves automatically by logical, ontology-based referencing. The context of dedicated information is (in sum) further represented by dynamic graphs which illustrate even complex relations among entries and their specific relation (properties). Illustrating the context of an object in a visually comprehensive way reduces complexity. Furthermore the history-function of Semantic MediaWiki allows to automatically keep track of changes.

Applying CIDOC-CRM in a Semantic MediaWiki environment provides meaningful information, regarding the ontology, which evolves during the course of usage. The CRM-Wiki is capable of documenting complex, open and even changing objects and could be easily linked to different net-based content provider, by using the wiki linkage or the RDF-export options.

Last but not least the integration of the browse-option (*Figure 4*) and automatically generate .dot-graphs illustrate the complex relation between different classes.



Browse wiki

D9 Sintel Master

L56 has pixel width	4096 +
L57 has pixel height	1744 +
L58 has thumbnail	File:D39 Sintel.png +
M1 codec	Tiff16 stream +
M1 frames	21312 +
Has query	D9 Sintel Master + , D9 Sintel Master + , D9 Sintel Master + , D9 Sintel Master + , D9 Sintel Master + , D9 Sintel Master +
Categories	D9 Data Object +
Modification date	15 March 2014 20:47:57 +

hide properties that link here

D10 Sintel Master +	L11 had output
D13 Sintel Master +	L19 stores
D10 Sintel Full HD +	L2 used as source

Figure 4. *Semantic Browsing*

Conclusion

Sharing information in a common system environment, where professionals of different cultural backgrounds such as artists, curators, archivists, conservators and theorists can meet, will become more and more important in the future. This means that a clearly structure but potentially extensible network of semantically tagged references and relations is required. It should be capable of keeping track of a whole set of thematically different documentation actions and objects – digital objects can easily become technically diverse (threat of technical obsolescence). Semantic relations to specific contexts and habitual modes such as working methods, ways of thinking and classifying/structuring/ storing information need to be tracked right at the production level in order to stay comprehensive. For that reasons the present case study implemented the CIDOC CRM concept to a Semantic MediaWiki. As planning and research tool the presented approach seems reasonable. Interested can get the CRM-Wiki import software directly from the authors.

References

[1] CIDOC CRM Special Interest Group & ICOM - International Council of Museums. (2013). Definition of the CIDOC Conceptual Reference Model: Version 5.1.2. Retrieved from http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.2.pdf

[2] Constantopoulos, P. & Dritsou, V. (2006). A CIDOC CRM – compatible metadata model for digital preservation. Retrieved from http://www.cidoc-crm.org/workshops/heraklion_october_2006/constantopoulos.pdf

[3] Doerr, M. (2003). The CIDOC CRM, an Ontological Approach to Schema Heterogeneity: Dagstuhl Seminar Proceedings 04391. Semantic Interoperability and Integration. Retrieved from http://drops.dagstuhl.de/volltexte/2005/35/pdf/04391.DoerrMartin.ExtAbstract.35.pdf

[4] Doerr, M. & Theodoridou, M. (2013). CRM dig: A generic digital provenance model for scientific observation Institute of Computer Science,.

[5] FORTH-ICS. (2013). CRMdig : An Extension of CIDOC-CRM to support provenance metadata. Retrieved from http://www.ics.forth.gr/isl/CRMext/CRMdig/docs/CRMdig3.0.ppt

[6] Blender Foundation (Producer). & Levy, C. (Director). (2010). Sintel. Digital Film Edition, USA. Retrieved from http://www.sintel.org/

[7] Lialina, O. (2013). Summer. Internetbased Artwork. Retrieved from http://art.teleportacia.org/olia/summer/

[8] Olensky, M. (2010). Semantic interoperability in Europeana. An examination of CIDOC CRM in digital cultural heritage documentation. (Dissertation). Humboldt-Universität zu Berlin, Berlin. Retrieved from http://www.ieee-tcdl.org/Bulletin/v6n2/Olensky/olensky.html

[9] Semantic MediaWiki. (2014). Main Page. Retrieved from http://semantic-mediawiki.org/wiki/Semantic_MediaWiki

Author Biography

Since 2012 Juergen Enge is head of the Centre for Information, Media and Technology at HAWK. Between 2006 and 2012 he directed the research field "Digital memory" at Karlsruhe University of the Arts. In numerous case studies, research and EU projects he investigated issues of preservation of complex digital objects. Since 2013 he heads the computer science curriculum of the graduate program MAS Preservation of Digital Art & Cultural Heritage (PDACH) at BUA.

Tabea Lurk is art historian and professor for digital conservation at Bern University of the Arts (BUA). Since 2008 she established the ArtLab of the Department of Conservation and Restoration at BUA. In 2011/12 she formed the graduate program Master of Applied Sciences BFH in Preservation of Digital Art & Cultural Heritage. In an honorary appointment, she is president of the special interest group on Digital Heritage of the Swiss Informatics Society.